Fermi National Accelerator Laboratory

FERMILAB-Conf-90/70

Recent Experiences and Future Expectations in Data Storage Technology*

Jack Pfister
Fermi National Accelerator Laboratory
P.O. Box 500
Batavia, Illinois 60510

April 1990

* Talk presented at the 1990 Conference on Computing in High Energy Physics, Santa Fe, New Mexico, April 9-13, 1990.

Operated by Universities Research Association Inc. under contract with the United States Department of Energy

"Recent Experiences and Future Expectations in Data Storage Technology"

Jack Pfister
Fermi National Accelerator Laboratory, Batavia, IL 60510

ABSTRACT

For more than 10 years the conventional media for High Energy Physics has been 9 track magnetic tape in various densities. More recently, especially in Europe, the IBM 3480 technology has been adopted while in the United States, especially at Fermilab, 8mm is being used by the largest experiments as a primary recording media and where possible they are using 8mm for the production, analysis and distribution of data summary tapes. VHS and Digital Audio tape have recurrently appeared but seem to serve primarily as a back-up storage media.

The reasons for what appear to be a radical departure are many. Economics (media and controllers are inexpensive), form factor (two gigabytes per shirt pocket), and convenience (fewer mounts/dismounts per minute) are dominant among the reasons.

The traditional data media suppliers seem to have been content to evolve the traditional media at their own pace with only modest enhancements primarily in "value engineering" of extant products. Meanwhile, start-up companies providing small system and workstations sought other media both to reduce the price of their offerings and respond to the real need of lower cost back-up for lower cost systems. This happening in a market context where traditional computer systems vendors were leaving the tape market altogether or shifting to "3480" technology which has certainly created a climate for reconsideration and change. The newest data storage products, in most cases, are not coming from the technologies developed by the computing industry but by the audio and video industry. Just where these flopticals, opticals, 19 mm tape and the new underlying technologies, such as, "digital paper" may fit in the HEP computing requirement picture will be reviewed. What these technologies do for and to HEP will be discussed along with some suggestions for a methodology for tracking and evaluating extant and emerging technologies.

INTRODUCTION

In any discussion of the elements of computing, it is obligatory to set a context. First, the computing in high energy physics has some distinguishing data characteristics:

a. Large data volume (1.0-10+TBytes)
b. Large record sizes (0.01-1.0MByte)
c. Variable computational intensity vs. input/output
d. Raw or summary data may have wide distribution for analysis

Second, the Model of Computing in HEP varies from laboratory to laboratory whether university, national or international in character. Understanding the relationship of the laboratory, where the data is recorded, to the collaborators capabilities locally and at their home institutions is crucial to success in the "Data Experience". Just how much data, how much computation - done where? over what time period? are well known issues, but not always well accounted for in developing the computing model and choosing the media and technology. Understanding the model is crucial to the next aspect - the evaluation model.

The third item in such a discussion must be the choice of an evaluation model and a technical tracking and assessment methodology. Formally or informally, they must exist unless you accept chaos as a default and an acceptable choice. The evaluation model used here is of my own making which has the bias of scars developed over the years and many opportunities to choose(or guess). The elements include:

1. assumptions about the evaluation; context and concept of operation
2. summary of functional requirements
3. throughput analysis - benchmarked if possible
4. capacity - benchmarked if possible
5. data interchange - compatibility, tested
6. robotic opportunities - availability, relevance
7. level of integration (inherent for the device to function in your environment)
8. robustness - failure and recovery rates
9. price and cost (they aren't the same!) - Look at the life cycle cost
10. standards - international, national, industry, local
11. standardization - media, form factor
12. source of supply; competition, availability, service
    - media
    - device
13. qualitative issues - quality assurance on media, device
14. distribution channels - how readily available
15. connectivity - industry vs. national & interpretation

All this would seem to a bit much after all you might say, "I just want to write my data". Though it looks daunting, I contend that using the foregoing model, we can easily qualify a vendor, a device, or a type of media without resorting to hysteria and hyperbole. It won't be as much fun but ....

TECHNOLOGY, CURRENT EXPERIENCE:

The experience with current technologies, for the most part, is well understood. None the less, each deserves some comment on it's viability and near term future, as well as "packaging" options and the extent to which robotics are employed.

NINE TRACK TAPE (1/2" wide x 2400 feet):

The mainstay in commercial and scientific computer centers for more than 20 years this media has evolved from a 2 MB capacity to a capacity of about 160 megabytes in standard format speeds in start/stop mode of 200 inches/second and transfer rates over one MByte/second[1]. It has improved over the years in reliability, has good archival retention characteristics, and if stored under proper conditions (temperature, humidity and air quality), is good for up to thirty years. However, the basic form factor 10.5 in. reels are cumbersome and bulky and require huge storage facilities at the HEP sites. The 10.5" reel flirted briefly with robotics the most successful of which was the Calcomp (which became in succession, Braegan, TRW, and now Sorbus) Automatic Tape Library which was still being built as late as 1985 and is still installed in a few U.S. National Laboratories including Fermilab. This media will be around for many years to come due to the numbers of these tapes written and their archival characteristics. It is worth noting, however, that, almost no major computer system vendor builds any nine track tape drives having left it to the third party market to provide this capability. The third party market is building drives largely in the middle range of performance - primarily for data exchange though high performance drives are still manufactured and available[2].

CARTRIDGE TAPE:

IBM 3480/3490 (1/2" wide x 600 feet):

"3480" and compatible cartridge tape is replacing the nine track tape in many computing centers. This device has been available for over five years and offers higher capacity than nine track (200 MB or ~35%) per unit of media in at least 1/5 the physical storage of nine track tape. For the computation center, it offers advantages in reduced costs of power, cooling and floor space (the critical advantage). This technology was better matched to the requirements for faster data transfer (initially 3.0 and recently up to 4.5 MB/second). The media format is 18 parallel tracks recorded at ~40k bits per inch (BPI) on 0.5 in. wide tape in a 4" x 5" cartridge.[3] The form factor (the unit of media) is quite conducive to robotics and Storage Technology Corporation (STK) and Memorex offer robotic devices. The STK device manages the media in units of 5000/silo (about 1 terabyte) and can be interconnected (cartridges passed from silo to silo) to a staggering 64 terabytes of "Nearline"[4] (as opposed to on-line or off-line).

This 18 track technology recently received a "midlife" kicker when the industry accepted a data compression algorithm which on a sample from the HEP experiments improved capacity up to 1.8. The big customers of this technology eagerly await a double density, double speed (with backward compatible) version which with compression could provide the most reliable 1GByte tape available in the industry. It is to be expected that when IBM announces availability STK will follow "within hours". This technology is also available in rack mount versions with DEC and SCSI interfaces. All versions have stackers holding 6-10 cartridges. Not all vendors have implemented a standard compression algorithm, so compatibility can be an issue.

At Fermilab, the "3480" technology was used on the Amdahl piece of the computing capability initially as backup and archival of data sets but now as a staging media for multiply accessed data that cannot justify permanent disk storage but is mounted many times a day but for relatively short periods of time. The utility of the media in the VAX cluster is not immediately obvious except for the convenience of exchange and a tribute to consistency.

8MM:

The media we hate to love but is so endearing because it currently has the best capacity (G Byte/shirtpocket/$ ratio) and price - a controller/drive can be purchased at half the cost of other more traditional media devices (<$5.0k). Unlike the serial media discussed so far, the 8mm is a very "modular" product built from Sony's 8mm "camcorder engine". The drive mechanism is packaged in the U.S. exclusively by Exabyte Corporation and sold to so-called value add resellers (VAR's) who (in general) add value by enhancing the product with various interfaces (QBUS, Unibus, IBM Channel), features (implementation of ANSI standards, search, controller/drive commands-search, unload, etc.) package with menu driven software product for backup and provide out of warranty service. They either sell directly or through a distributor to the computing public. This presents the owner/operator of these devices with a nearly endless list of possible opportunities for mischief as the product finds it's way into production usage. If your application requires multiple interfaces (QBUS, Unibus) the odds are high that if different suppliers are involved that the implementation of standards or features are different. Technically, the 8mm uses helical scan recording as a series of parallel tracks 2 " long at an acute angle to the edge of the tape. The density is somewhat greater than "3480" at 43k bits/inch and a track density of about 800/inch. Actual transfer rates vary widely and are highly dependent on the VAR's handywork but we have in a variety of application seen 80-180 KB/s. The best speed and therefore, transfer that can be achieved is at the beginning governed by the tape movement and head movement which is current models is 0.5 in./s and 150 in./s respectively. It is clear that any improvement in tape movement in the future has big gain potential.

The Exabyte Company has announced that in the June-July (1990) period a double density (5 GB) double speed version, the EXB-8500 product will be available with full functionality including backward compatibility with the current product by the late fall (1990). Thus, the 8mm achieves the respectable transfer rate of the mid-range 9 track tape drives (300-500 kBS) but with by any standard - an incredible 5 gigabyte capacity.

A detailed technical assessment of this product is difficult because there is not, beyond Exabyte, a constant standard or single application. The device has been shown to be fairly robust as it left Exabyte. Beyond that, there is a separate story (good and bad) for each vendor (VAR) and application that we have installed at Fermilab. These issues raised here and subsequently show why a methodology and good bookkeeping are essential.

Performance and capacity tests are or can be determined objectively. Data interchange on compatibility between different vendors should also be tested appropriate to the site and other collaborators machines. The functionality needed and supported should be tested. At Fermilab, in May 1989, three vendors claimed DEC HSC and VMS copy compatibility and indeed nearly one year later we could only substantiate one claim. The media itself has become an issue "Hi-Metal", "Hi-Video", "Video Quality", "Data Quality" - price ranges factor of five, qualitative issues - in some cases zero difference under test conditions. Product information as to stability, performance and shelf life are not obtainable from media suppliers.

The number of modules and nodules that permit this product to function at all is so high that it must receive the Level of Integration Complexity Award. See Figure 1, for a somewhat simplified view of the 8mm device as it moves through the "integration" process.

Robotics are being made available for the 8mm devices in the form of carrousels which are deliverable today from Summus - 54 cartridges with two drives and soon from Exabyte with 120 cartridges and four drives. Also to be expected is a stacker product from Exabyte which will handle 10 cassettes is a portable tray. Fermilab has developed a stacker which holds 12 cassettes and can be equipped with a scanner which is under consideration for commercialization. The 8mm products for all the integration complexity have reached all corners of the computing at Fermilab and all the major systems and workstations have some interface supported. Systems currently configured include:

DEC - HSC, QBUS, UNIBUS; Silicon Graphics - SCSI,
Amdahl - IBM Channel; Sun - SCSI; ACP - UNIBUS and VME

This technology would seem to have a lot going for it. The transfer rate improvement though welcome and long awaited needs another factor of 2-4 improvement to better match the throughput potential of current main-frame class, the newer workstations and servers.

VHS:

The VHS cassette and drives used for data have been available for the past three years as a backup device for personal computer and small system products. It has not, however, enjoyed the same popularity as the 8mm over the same period. Another product using VHS and targeted for higher capacity and throughput comes from Honeywell. This product known as the Very Large Data Store (VLDS) has a per cassette capacity of 5.2 Giga Bytes and with its SCSI controller, the VLDS supports asynchronous transfers (up to at 1.0 MB/sec sustained), Synchronous (up to 4.0 MB/sec) and high speed data port (up to 10.0 MB/sec burst).

This helical scan device records data at a 6^{o} angle which permits a 4" long path in which 16 KBytes of user data is recorded. Two paths provide a 32 KB fixed recording block. These blocks are assigned a block number which can be used to recall data via a controller and generate a directory. Searches are conducted in a "fast forward mode". There are four horizontal track (parallel to the tape edge) two top and two bottom identified respectively as direct channel, file mark, control and servo pulses. The head rotates at 60 rps for an

effective head to tape speed of just over 450 inches per second. The best known robotics, Honeywell's VLA/VLAS, Very Large Archive/Very Large Archive server provide a capacity of just over 3 TB (600 cassettes) and can be equipped with up to 5 VLDS drives. Access to any cassette and load/unload can in the worst case take 24 seconds. Two rotating drums of 300 cassettes are used to present a stack to the cassette handler, a bar code reader is provided. This product has a high degree of integration and includes capabilities for matching internal and external labels, network file service using TCP/IP and has application support for FTP, RCP and TELNET. Data management services are provided through a relational data base (which has user and system level queries) which can track 65K physical volumes and 90 million files, permits dead space recovery, and supports import/export of files up to 4 GB. The VLA can appear as a standard node in server networks. To date, Fermilab has not evaluated the product although one Fermilab experiment is using the VLDS in a data acquisition application. Their data is subsequently converted to 8mm.

DAT:

Digital audio tape is another technology which employs the helical scan technique but on 4mm tape packaged in a credit card size cassette with capacities up to 1.2 Giga Bytes. Primarily targeted as a backup device, DAT is marketed for personal computer (IBM and Macintosh) and workstations (SUN, DEC) and has interface support for SCSI, PERTEC and Qic-02. The recording is at just over 6° relative to the edge with a linear recording density of 60K bpi. The read/write drum rotates at 2000 rpm and the tape moves at just under a third of an inch per second, yielding an effective tape speed of just over 120 ips and a nominal transfer rate of 190 KBS. It's principal advantage over 8mm in backup/restore use is the search speed to recover files. Fermilab has installed one device for evaluation in one of the data center ACP systems for evaluation and production use as a backup device. The results thus far are very good in this application. Fermilab has not yet done timing and transfer speed tests to compare with vendor claims. Neither robotics nor stackers have been announced to date but are in vendor plans. One concern in DAT is the existence of two "standards" for recording: RDAT (supported by DEC, HP, SUN) and DDAT (supported by Apple, Hitachi and Mitsushimi). Double length and double density for a 4 gigabyte capacity is anticipated. This is a prime case for a wait and see before committing wholesale.

MAGNETIC DISK:

The magnetic disk technology now over 30 years old has gone from kilobytes to gigabytes and access measured in seconds to ten milleseconds, and provides random access to relatively small blocks of data typically 512 bytes. This is a technology which continues to improve it's areal density and therefore cost, reliability and performance. Areal densities today exceed 100 million bits/sq. in. Long the mainstay in computing centers this technology is now very affordable for even the low-end workstation system. There are pressures at the high end of the performance curve to provide better immediate access performance for super systems with 10+ MB/sec

channels speeds. This technology is well understood, has a very well disciplined approach for improvement in a highly competitive market. The interface standards too are well disciplined. This bodes well for the full range of customer requirements for page/swap, direct access, random access and as intermediate storage such as a staging place for short term use by serial media or a copy point for subsequent file archival. The most recent "innovations" are caching and using the multiple spindles in various geometries including striping to further enhance performance. These enhancements initially the perview of the high-end system are becoming available at the workstation level with SCSI interfaces and transfer rates at the 3-4 MB/sec.

EMERGING TECHNOLOGIES

To say what is an emerging as technology as opposed to the application of a technology to new issues as in high energy physics is problematic. The reader should consider the division as arbitrary but not capricious.

OPTICAL DISK TECHNOLOGY:

The optical technologies are often considered as complementary to magnetic disks. They offer higher density and can be accessed in serial or random modes, albeit at lower data transfer rates. Three types of optical technology are: **CDROM** (compact disk read-only memory) which uses a master to replicate the same data to many disks by an injection molding process. The typical applications are software distribution and data, historical or encyclopaedic in character. Capacity is rated at 600 MB and access times are up to 500 millesconds. Subsystems supporting SCSI are available.

WORM: (Write-once, read-many) which has data written by a host computer but cannot be re-written without destroying the original data. The nature of the media, usually plastic encased, makes it ideal for archival applications especially in a hierarchically managed storage systems. Worm drives are available in 3.5 in. to 14 in. form factors with capacities from 0.5 to 2 gigabytes per platter. Sustained transfer rates range from a low 30 (write) to a high of 400 (read) KBS. Juke boxes in the one to three hundred gigabyte capacity are available with a number of standard interfaces supported. A feature of some implementations is the ability to use the devices either in random or serial access modes.

ERASABLE OPTICAL: Is the newest in this category and has functional characteristics like magnetic disk but with lower costs for a medium performance removable storage. Computing centers may find this media useful in cyclic applications such as incremental and full backup applications. The trade-off is in lower performance for a very large online data capability. The erasable opticals employ three recording techniques with magneto optical the leading favorite. It has capacities in the 5.25" form factor of nearly 0.6 GB and 1.0 - 2.0 in the 12" form factor. Host interfaces exists for SCSI and support user data transfer of rates at up to 680 KBS. Actual capacity and data rates are dependent upon what density is chosen for blocks per sector. The drives are low cost but the media currently is about $250.

Jukeboxes are available which have two drives and hold 56 platters. The media is reckoned to have a ten year life. It is to be noted that it is very early days for this technology, standards and manufacturing techniques are still undergoing refinement. A dramatic improvement in media cost could make this an attractive option in high volume data applications.

SONY DIGITAL INSTRUMENT RECORDER:

The DIR-1000 is the latest product from Sony Corporation and uses 19mm type D-1 broadcast standard tape in large, medium and small cassettes with user data capacities of nearly 100, 40 and 12 Gigabytes respectively. Aside from the drives ability to automatically adjust to different sized cassettes, it also has a variable data rate from just over 1 MB/sec to 32 MB/sec in one of six selected steps, i.e., 1, 2, 4, 8, 16, 32 MB/S. Bit error rates (corrected) in the 1x10E-10 are claimed. This device too employs helical scan recording techniques writing a track set nearly 6.7 inches long on a 5.4^{o} angle to the edge. Three control and annotation tracks are written parallel to the tape edge. Currently available, this device has a VME interface and additional interfaces SCSI II and HPPI IBM standard are under development. The commercial list price is $250K in quantity one. To date, one unit has shipped for an imaging application. We will undoubtedly be hearing more about this product as it has excellent prospects for high data rate and data volume applications.

"DIGITAL PAPER" TECHNOLOGY (OPTICAL TAPE):

In the Spring of 1988, news and papers began appearing heralding the arrival of "digital paper" from ICI a chemical giant in Europe and represented by it's subsidiary ICI Imagedata in the U.S. In late November at Comdex, this worm technology was said to in a "reel of the same proportions as a conventional 10.5 in magnetic tape ... store 600 gigabytes of data". By the following Spring, it is up to a terabyte on a reel of tape and a gigabyte on a floppy diskette. Two companies, BOSCO, a subsidiary of Iomega (the floppy disk) and CREO Products of Canada (a 35mm tape) had begun product development of this write once technology. The so-called "digital paper" is based on a polyester substrate (Melinex) with a sputtered layer of metal to make it reflective and then covered with an ICI developed polymer dye formulated to absorb specific wave lengths. The polymer dye is deformed by the head in the writing process. The metal layer dissipates heat from the focus point requiring less laser energy for creating the depression.[6]

By March 1990, the floppy disk is off the table as IOMEGA has stopped work because, as the article states, the drive technology is complete but the media is not yet available in volume and "the company saw other opportunities for a better return on investment".[7] The same article goes on to stat that ICI still intends to license the technology and feels "we proved the technology ... the media is still evolving ... there is still work to be done."[8]

The optical tape work, however, continues by CREO on their Model 1003 drive. The data is formatted as 4 byte-wide words (32 bits) in 20KB or 80KB blocks. The tape is nearly 2900 feet (880m) in length. Data transfer rates are said to be at 3 MB/S.[9] An end to

end search would take 60 seconds. Speculation abounds as to the products availability in late 1990. The price of the drive and media are expected to be $200K and $10K respectively.[10]

SUMMARY AND CONCLUSIONS

This paper has attempted to survey current and future opportunities of storage devices possibly relevant to HEP (Table 1) and attempts to provide a structure for tracking and evaluating various technologies. There is no intentional attempt to declare winners or losers among a set of technologies but to provide a range of apparent options which could be evaluated for appropriateness in each HEP setting. There are some obvious trends among them. The shortening technical life of products means we will have to retain more of the older technologies to read the data that is not migrated or is incompatable with new technologoes. The level of integration from the media through the application is more difuse and the ownership of and ability to deal with problems by the customer and maintenance organizations is more difficult. Hetrogeneous computing environments will have to develop strategies for compatability across computing platforms and the ubiquity of operating system upgrades. Greater depth and breadth in configuration control and management are required. Finally, one can get into a technology too early - what ever it's potential, the opportunity (with apologies to Pogo) may be insurmountable.

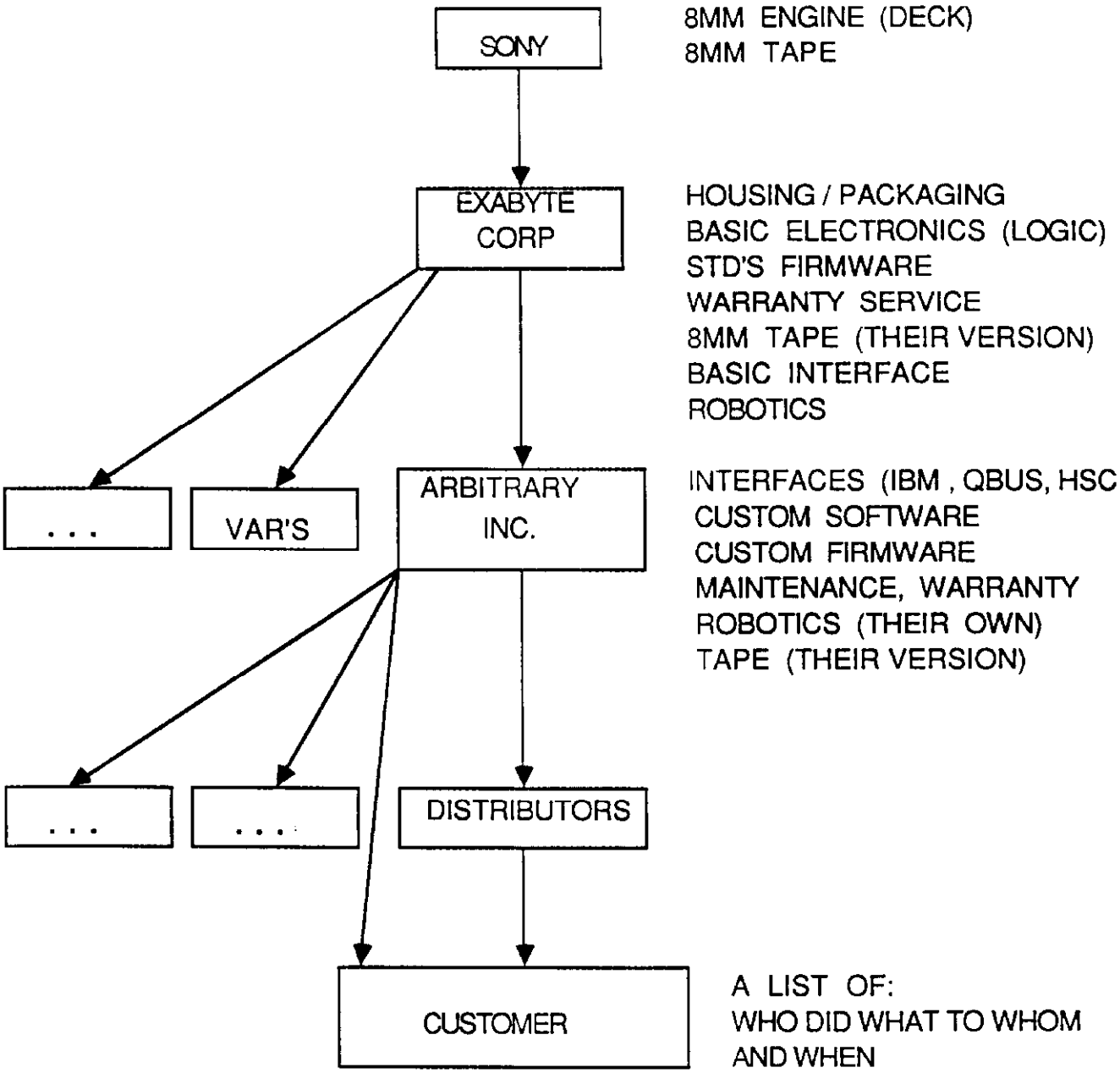

FIGURE 1.
8MM INTEGRATION PROCESS

TABLE I

DATA STORAGE TECHNOLOGY
(by capacity)

MEDIA (TYPE)	CAPACITY (MB)	DATARATE (MB/S)
NINE TRACK TAPE	160	0.3 - 1.5
"3480"	200	3.0 - 4.5
CDROM	600	0.15
ERASABLE OPTICAL	600	0.6 - 0.7
DAT-4MM	1000	0.18 - 0.25
DIGITAL PAPER-FLOPPY	1000	1.1
8MM	2500	.08 - 0.25
MAGNETIC DISK (HIGHEND)	2700	3.0 - 4.5
OPTICAL DISK	3200	0.12 - 0.35
VHS (HONEYWELL)	5200	2.0 - 4.0
19MM (SONY)	100,000	1.3 - 32.0
DIGITAL PAPER - TAPE	1,000,000	3.0

REFERENCES

1. Digital Storage Technology Handbook, Digital Equipment Corporation, 1989 P. 6-1
   Note: This is a generally useful and readable book for a general understanding of various computing technologies.

2. Data Sources Hardware, 1st Edition, Vol. 1, Zitt-Davis, N.Y., N.Y. (1990)

3. Digital Storage Technology Handbook P.6-25.

4. "Nearline" is a trademark of Storage Technology Corporation.

5. Digital Storage Technology Handbook, P. 6-11, 12

6. Tom Williams, "Computer Design", Vol. 28, NR 7 (April 1, 1989)

7. Brian Deagon, "Electronic News", P. 17 (March 5, 1990)

8. IBID

9. Tom Williams, "Computer Design"

10. Doug Chandler, "PC Week", Vol. 5, NR 33 (August 15, 1988)

JOP:bf